Case Study: ENVRI Science Demonstrators with D4Science

Leonardo Candela[1(✉)], Markus Stocker[2,3], Ingemar Häggström[4], Carl-Fredrik Enell[4], Domenico Vitale[5], Dario Papale[5], Baptiste Grenier[6], Yin Chen[6], and Matthias Obst[7]

[1] National Research Council of Italy, Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo", Via G. Moruzzi, 1, 56124 Pisa, Italy
leonardo.candela@isti.cnr.it
[2] TIB Leibniz Information Centre for Science and Technology, Welfengarten 1 B, 30167 Hannover, Germany
markus.stocker@tib.eu
[3] MARUM Center for Marine Environmental Sciences, PANGAEA Data Publisher for Earth and Environmental Science, Leobener Strasse 8, 28359 Bremen, Germany
[4] EISCAT Scientific Association, Box 812, 981 28 Kiruna, Sweden
{ingemar.haggstrom,carl-fredrik.enell}@eiscat.se
[5] Department for Innovation in Biological, Agro-Food and Forest Systems (DIBAF), University of Tuscia, Via San Camillo de Lellis, 01100 Viterbo, Italy
{domvit,darpap}@unitus.it
[6] EGI Foundation, Science Park 140, Amsterdam, The Netherlands
{baptiste.grenier,yin.chen}@egi.eu
[7] Swedish Lifewatch, Gothenburg, Sweden
matthias.obst@marine.gu.se

Abstract. Whenever a community of practice starts developing an IT solution for its use case(s) it has to face the issue of carefully selecting "the platform" to use. Such a platform should match the requirements and the overall settings resulting from the specific application context (including legacy technologies and solutions to be integrated and reused, costs of adoption and operation, easiness in acquiring skills and competencies). There is no one-size-fits-all solution that is suitable for all application context, and this is particularly true for scientific communities and their cases because of the wide heterogeneity characterising them. However, there is a large consensus that solutions from scratch are inefficient and services that facilitate the development and maintenance of scientific community-specific solutions do exist. This chapter describes how a set of diverse communities of practice efficiently developed their science demonstrators (on analysing and producing user-defined atmosphere data products, greenhouse gases fluxes, particle formation, mosquito diseases) by leveraging the services offered by the D4Science infrastructure. It shows that the D4Science design decisions aiming at streamlining implementations are effective. The chapter discusses the added value injected in the science demonstrators and resulting from the reuse of D4Science services, especially regarding Open Science practices and overall quality of service.

Z. Zhao and M. Hellström (Eds.): Towards Interoperable Research Infrastructures for Environmental and Earth Sciences, LNCS 12003, pp. 307–323, 2020.
https://doi.org/10.1007/978-3-030-52829-4_17

Keywords: Virtual research environment · Open science · D4science · Science demonstrators · Science communities

1 Introduction

Science is highly digital, collaborative and multidisciplinary and science practices have been changed in recent decades [6]. These changes are induced by the opportunities offered by the developments in information technologies and infrastructures and are impacting the whole research lifecycle – from data collection and curation to analysis, visualisation and publishing. Research communities are dynamically aggregated, working environments conceived to support research tasks are virtual, heterogeneous and networked across the boundaries of research performing organisations. Scientists are thus asking for integrated environments providing themselves with seamless access to data, software, services and computing resources they need in performing their research activities independently of organisational and technical barriers [5]. In these settings, approaches based on ad-hoc and "from scratch" development of the envisaged supporting environments are neither viable (e.g. high "time to market") nor sustainable (e.g. technological obsolescence risk).

Environmental science is not eluding these changes, rather it is fully affected by them. A rich array of environmental research infrastructures is being organised and developed to provide their designated communities with computing resources, services and facilities for data collection and collation, processing, analytics, and publishing. Initiatives such as ENVRIplus [1] and the European Open Science Cloud [10] have been launched to make available state-of-the-art solutions for data management thus making the development and operation of environmental research infrastructures more efficient.

In spite of these developments, researchers and scientists are still struggling with the lack of working environments tailored to their specific needs, especially when operating in multidisciplinary contexts.

In this chapter, we present the D4Science-based solution for developing and operating *virtual research environments* for different communities of practice identified in selected science demonstrators. D4Science [2, 3] enacted virtual research environments promote the re-use of domain-specific existing data and services, the co-creation and co-development of the envisaged working environment, and the use of state-of-the-art solutions for collaboration, communication and Open Science.

This chapter presents four concrete and diverse science demonstrators. These cases concern (a) providing scientists willing to analyse data collected by EISCAT radars with a collaborative working environment, (b) implementing shared, standardised and reproducible data processing and quality control (QC) procedures for long-term eddy covariance (EC) flux datasets, (c) providing scientists involved in atmospheric new particle formation event analysis with computational environments for event identification and classification with built-in analysis (derivative) data FAIRification, and (d) providing scientists seeking to increase our knowledge of biodiversity organisation and ecosystem functions with a working environment to test models. In particular, the challenges and the resulting prototypical working solutions (with their pros and cons) community of practices managed to develop are presented and discussed.

2 The Collaborative Working Environment for Data Analysis

EISCAT_3D will differ from other environmental research infrastructures with respect to its configurability and data volumes. A typical environmental RI measures well-defined parameters and stores the data in a specified way. EISCAT_3D, on the other hand, will be a flexible, multi-purpose instrument. Archived data can be reanalysed to extract parameters in complementary research domains, typically for example both electron and ion densities and temperatures in the ionosphere and the influx of meteors from space. Data access rules also apply according to the agreement between EISCAT members, including embargo times for PIs of experiments. This means that users must be allowed to upload and run their own analysis software on archived EISCAT_3D datasets to which they are granted access.

The use of big data and supercomputing systems will be unfamiliar to typical EISCAT_3D scientists, so authentication, search and analysis should be handled by a portal. This portal should have a search GUI as well as APIs for script-based access. This line of work is also further developed by EGI and in the European Open Science Cloud Hub Competence Centre (EOSC-Hub CC) for EISCAT_3D.

In the framework of ENVRIplus, EISCAT_3D has been a pilot case in using D4Science. An advantage of the D4Science portal is that it allows uploading user software in many languages. The online R studio is well developed, but GNU/Octave, Python and several other languages are also available. Like in the DIRAC portal development in the EOSC-Hub CC, the science demonstrator had to work on existing data from the present EISCAT radars. The realtime graph plotting routine was selected as a common analysis case. This is Matlab software but runs also in Octave, which eliminates the need for a software license. File format conversion software written in Python with the HDF5 library has also been tested.

Figure 1 shows the D4Science file management GUI, which presents a familiar interface to the user. Here, program and data files were uploaded.

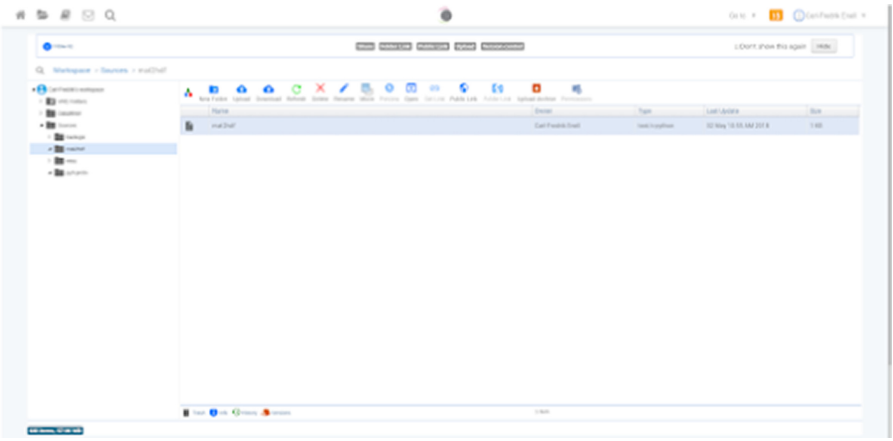

Fig. 1. The D4Science file GUI.

Figure 2 shows the DataMiner interface, where selected jobs are submitted for execution with selected input. Finally, Fig. 3 shows results being listed.

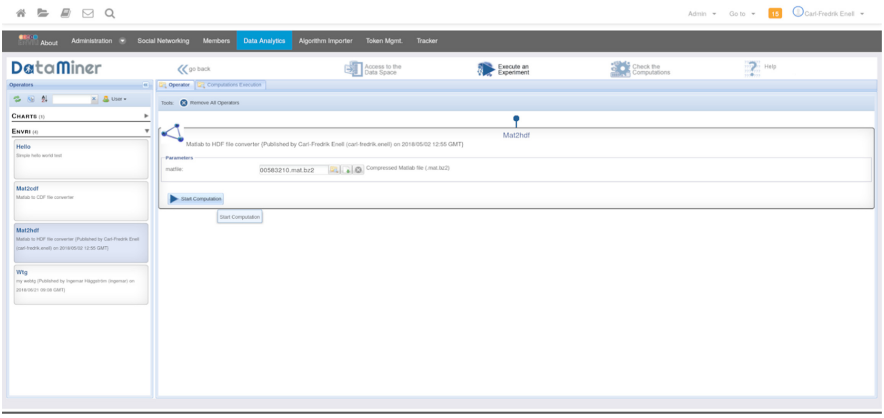

Fig. 2. The D4Science Data Miner job submission interface.

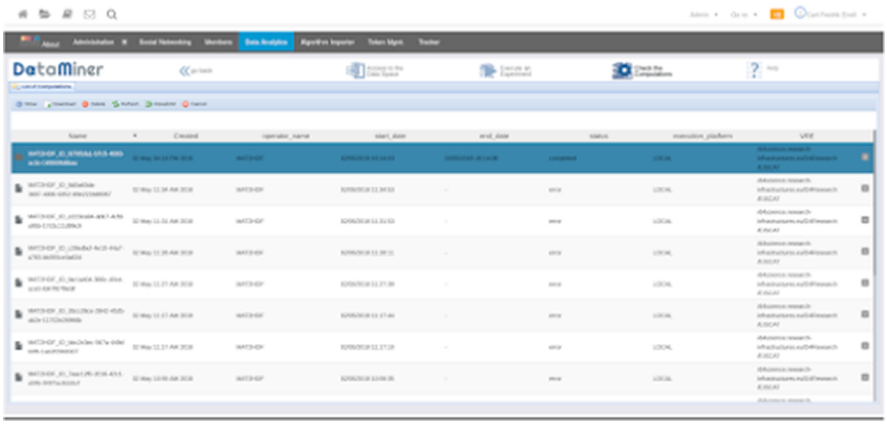

Fig. 3. The D4Science Data Miner list of completed jobs.

In EOSC CC, the development of the DIRAC system[1], originating from the LHCb detector at CERN, is ongoing. DIRAC is a job submission system that could benefit from a frontend such as D4Science. At this stage, EGI Check in AAI has been added to DIRAC to accommodate EISCAT access rules. The DIRAC metadata catalogue will also be extended for our purposes. Data will be stored using a replicating file management system such as Rucio, which is a development for the LHC ATLAS detector at CERN[2]. As for file storage backends, the project has been considering dCache or a parallel system such as Lustre or IBM GPFS.

[1] http://www.diracgrid.org.

[2] http://rucio.cern.ch.

Another pilot project was granted by EUDAT where EISCAT collaborated with CSC, Finland. Here, metadata constructed from EISCAT's existing Level 2 and 3 data systems were uploaded to B2SHARE. Figure 4 shows a sample B2SHARE entry. This would provide a common search interface for the two data levels. The existing data server of EISCAT has only basic search functions for listing the two levels together (namely the online schedule, ordered by date). We also foresee that all metadata will be provided for harvesting into B2FIND.

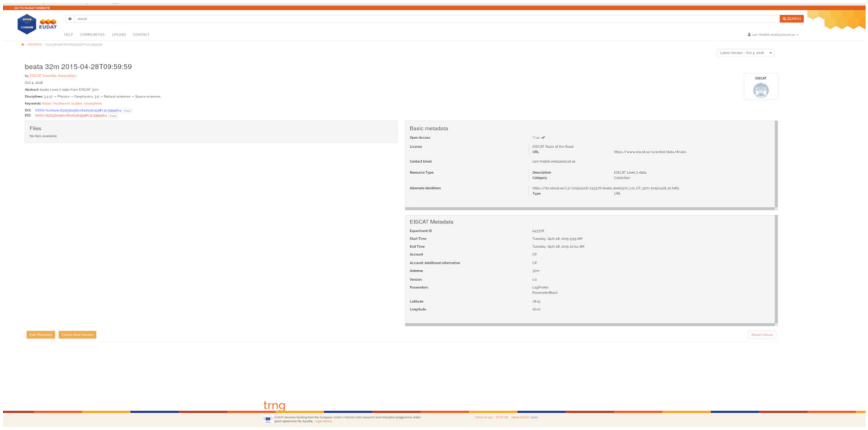

Fig. 4. A sample B2SHARE entry for existing EISCAT data, created using the B2SHARE REST API in Python software.

3 The Eddy Covariance of GHGs Fluxes Use Case

The eddy-covariance (EC) technique is considered the most direct and reliable method to calculate flux exchanges of the main greenhouse gases (GHG) over natural ecosystems and agricultural fields. The resulting measurements are extremely important to characterise ecosystem exchanges of carbon, water, energy and other trace gases and are widely used to validate or constrain parameter of land surface models via data assimilation techniques.

EC fluxes calculation involves a complex set of data processing steps that, beyond the knowledge of the technique, requires a considerable amount of computational resources. This might constitute a constraint for RIs (e.g. ICOS) that aim to simultaneously process large raw dataset sampled at multiple sites in Near Real Time (NRT) mode (i.e. provide each day fluxes estimates relative to the previous day).

The ambitious goal of this pilot investigation is to provide a computationally efficient tool able to process EC raw data and offer users the possibility to calculate fluxes according to the multiple processing scheme [9]. The ultimate aim is to establish a service that can be used by RIs that use this micrometeorological technique to measure exchanges of greenhouse gases and energy between terrestrial ecosystems and atmosphere (e.g. ICOS, LTER and ANAEE).

3.1 Virtual Research Environment

The EC technique involves high-frequency sampling (e.g. 10 or 20 Hz) of wind speed and scalar atmospheric concentration data and yields vertical turbulent fluxes. EC fluxes are computed within a finite averaging time (normally 30 min) from the covariance estimates between instantaneous deviations in vertical wind speed and gas concentration (e.g. CO_2) from their respective mean values, multiplied by the mean air density [4].

Despite the simplicity of this idea, a number of practical difficulties arise in transforming high-frequency data into reliable half-hourly flux measurements. To cope with these issues, here we used the tools implemented by the EddyPro® Fortran code [8]) an open source software application available for free download at https://www.licor.com/env/products/eddy_covariance/eddypro.html. The choice of EddyPro® software is motivated by *i*) the availability of different methods for data quality control and processing (e.g. coordinate rotation, time-series detrending, time lag determination, spectral corrections, and flux random uncertainty quantification), *ii*) the availability of the source code and *iii*) the fact that the software is based on a community-developed set of tools.

Required for the processing of EC raw data through EddyPro® software, are 1) the availability of metadata information about the EC system setup and raw data file structure, and 2) the choice of a suitable combination of processing options.

Concerning 1), users have to provide a standardised metadata file in.csv format (metadata.csv, see Table 1). This file constitutes the input of an R script that automatically builds the mandatory files ingested into the EddyPro® software (i.e. the *.metadata* and *.eddypro* files) developed ad hoc for this exercise. The organization and name of the metadata variables is based on an international standard (BADM) used also in the USA network AmeriFlux. The format of the.csv has been instead designed in order to develop a template easy to prepare by individual scientists and organised RIs.

It is important to note that in the current implementation only a few sensors are supported (the one used in ICOS) but the structure has been prepared in order to be ready to add new sensors and new processing methods, options and combinations.

In case of NRT data processing, in order to perform part of the flux corrections (i.e. spectral corrections and planar fit), 5 additional configuration files are needed: planar_fit.txt, spectral_assessment_badr.txt, spectral_assessment_lddr.txt, spectral_assessment_bapf.txt, spectral_assessment_ldpf.txt. They can be obtained by specific EddyPro® runs based on long periods of data (at least one month of data is usually required for a consistent parameter estimation). The above files have to be placed together with EC raw-data files in an archive folder (data.zip) which will constitute the input file of the current implementation (see Fig. 5).

The use of different processing options leads to different flux estimates. Discrepancies in flux estimates are caused by systematic errors introduced by the methods used in the raw-data processing stage. Since there are not tools to establish a priori which is the best combination of processing options providing unbiased flux estimates, the viable solution, proposed by [11] and implemented here, involves a multiple processing scheme where EC flux data are calculated according to different combinations of methods.

In particular, EC fluxes are calculated according to four different processing schemes resulting from a combination of block average (ba) or linear detrending (ld) and double rotation (dr) or planar fit [12] (pf) processing options (for details see [4]). All other

processing options remain unchanged: maximum cross-covariance method for time lag determination, spectral correction method proposed by Fratini et al. (2012), statistical tests by Vickers and Mahrt (1997) and by Foken and Wichura (1996) for data quality control, method by Finkelstein and Sims (2001) to random uncertainty quantification.

Table 1. Description of Metadata to provide in the metadata.csv file.

Column	Variable Label (File Header)	Description (Units)
1	SITEID	Official EC station code following the FLUXNET standards (CC-Xxx)
2	LATITUDE	Geographic latitude ([-90,90] from S to N, decimal)
3	LONGITUDE	Geographic longitude ([-180,180] from W to E, decimal)
4	ALTITUDE	The altitude of ecosystem under study (m)
5	CANOPY_HEIGHT	Distance between the ground and the top of the plant canopy (m)
6	SA_MANUFACTURER	Manufacturer of the sonic anemometer (currently only gill)
7	SA_MODEL	Model of the SA (currently only SA-Gill HS-50 or -100)
8	SA_SW_VERSION	The embedded software version of the SA
9	SA_WIND_DATA_FORMAT	The format of wind data (currently only uvw)
10	SA_NORTH_ALIGNEMENT	Specify whether the SA's axes are aligned to transducers (axis) or spars (spar)
11	SA_HEIGHT	The vertical distance between the ground and the centre of the device sampling volume (m)
12	SA_NORTH_OFFSET	Specify the SA's yaw offset with respect to local magnetic north (degree positive eastward)
13	GA_MANUFACTURER	Manufacturer of the gas analyser (currently only licor)
14	GA_MODEL	Model of the GA (currently only GA_CP-LI-COR LI7200)
15	GA_SW_VERSION	The embedded software version of the GA
16	GA_NORTHWARD_SEPARATION	The distance between the centre of the sample volume of the GA and the SA as measured horizontally along the north-south axis (cm)
17	GA_EASTWARD_SEPARATION	The distance between the centre of the sample volume of the GA and the SA as measured horizontally along the east-west axis (cm)
18	GA_VERTICAL_SEPARATION	The distance between the centre of the sample volume of the GA and the SA as measured along the vertical axis (cm)
19	GA_TUBE_DIAMETER	The inside diameter of the intake tube (mm)
20	GA_FLOWRATE	The flow rate of the intake tube (l/min)
21	GA_TUBE_LENGTH	The length of the intake tube (cm)
22	FILE_DURATION	The time span covered by each raw file (min)
23	ACQUISITION_FREQUENCY	The number of records per second in raw files (10 or 20 Hz)
24	FILE_FORMAT	Specify the format of raw files (ASCII or BIN)
25	FILE_EXTENSION	Specify the raw files extension (e.g..csv,.txt or.dat)
26	LN	Logger number (from 1 to 10)
27	FN	Number of the file generated by the logger (from 1 to 10)
28	EXTERNAL_TIMESTAMP	0 or 1 if the timestamp in the file name refers to the beginning or the end of the averaging period, respectively.
29	INTERNAL_TIMESTAMP	1 if there is a timestamp internal to raw files, otherwise 0.
30	EOL	Specify the end of the line of raw files (e.g. lf)
31	SEPARATOR	The character that separates individual values in raw files
32	MISSING_DATA_STRING	Specify the character string used for missing data in raw files (e.g. NA, NaN, -9999)
33	NROW_HEADER	The number of rows in the header of the raw file
33+1	COLNAMES_1	Variable name in the first column of the raw data file
33+j	COLNAMES_j	Variable name in the j-th column of the raw data file
33+N	COLNAMES_N	Variable name in the last column of the raw data file

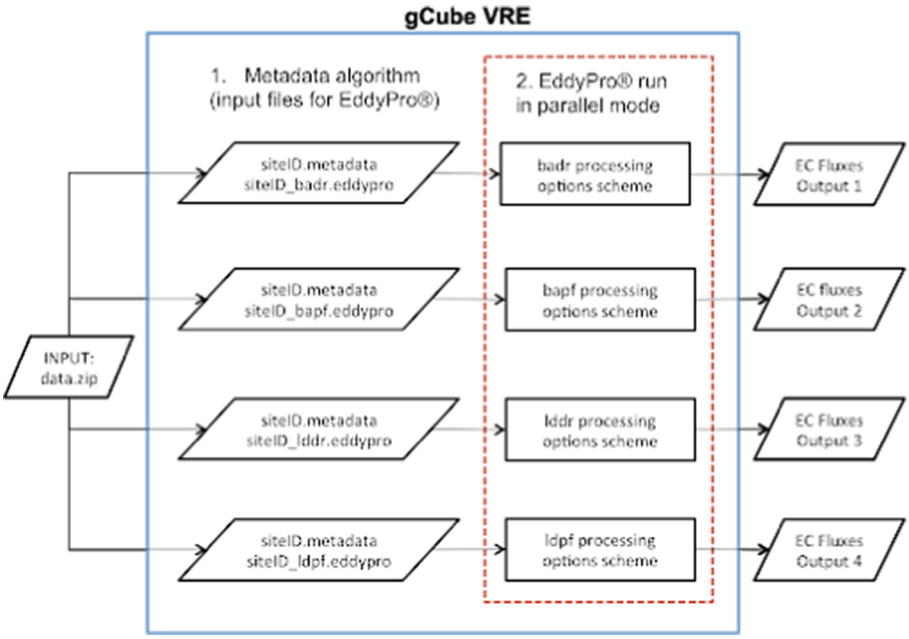

Fig. 5. EC data processing path.

To reduce the computational runtime, the implementation of the four processing schemes aforementioned is performed in parallel mode in the gCube Virtual Research Environment (VRE). The processing path is defined as in Fig. 5 (for an illustrative example see https://www.youtube.com/watch?v=ssHAfwXVF0A).

3.2 Benefits

The implementation of a multiple processing scheme as illustrated above and the direct management and use of metadata according to international standard in the eddy covariance community constitutes a novelty in the context of EC data analysis. The main advantage of multiple processing is twofold. On one hand, it offers the possibility of an extensive evaluation of the effect each method has on flux data estimation. On the other hand, by combining the output results as described by [11], it is possible to obtain more consistent estimates of the uncertainty associated with EC fluxes. The direct use of metadata instead ensures the needed flexibility for a large use of the tool if the new sensors are added in the system.

The efficiency of parallel computing implemented in the VRE drastically reduces the computational runtime required to obtain flux estimates from different processing options schemes. When using EC raw data from a single observation tower, the estimated computational time required for an NRT run is about 4 min, similar to those required for the run of a single processing scheme. This constitutes a clear advantage for any user and in particular, for RIs aiming at analyzing routinely large amounts of data. Although, here we selected only four processing option schemes, the efficiency of parallel

computing implemented in the VRE offers the possibility to increase the number of processing schemes suitable for the EC data processing and post-processing steps. This might considerably improve our understanding of the performance of methods developed for EC raw-data processing and the interpretation of resulting fluxes.

4 New Particle Formation Event Analysis

Atmospheric new particle formation (NPF) is a worldwide observed phenomenon that affects human respiratory health and the global climate [7]. NPF is studied by analysing (specifically, interpreting) the particle size distribution of polydisperse aerosol as measured by a differential mobility particle sizer at specific spatio-temporal locations (thereafter observational or primary data). The Finnish Station for Measuring Ecosystem-Atmosphere Relations[3] (SMEAR) research infrastructure operates such instruments at multiple spatial locations, including at Hyytiälä in southern Finland. The research infrastructure systematically publishes the collected observational data using SmartSMEAR. The observational data is thus accessible to researchers, worldwide. With the SmartSMEAR API, the data is also accessible programmatically.

To study NPF, atmospheric physicists analyse observational data to detect and characterise NPF events. During events, new particles initially form and then grow in diameter size, typically over the course of a few hours during the daytime. The detection of such events is typically performed manually by visualizing observational data for specific spatio-temporal locations (Fig. 6). Atmospheric physicists utilise such visual primary data products to determine whether or not an event occurred at the specific day and place. Events are then characterised (i.e., described) for their attributes, such as event start and end times, classification, or growth rate. With such primary data interpretation activity, atmospheric physicists generate derivative secondary data (here data about NPF events). Secondary data are subsequently used in statistical analysis, e.g. to compute descriptive statistics, thus resulting in derivative tertiary data which are sometimes published in the scholarly literature.

The FAIRification[4] of secondary and tertiary data is an important challenge for this research community, which consists of some hundreds of researchers organised in dozens of research groups (personal communication). While primary data are relatively FAIR, the derivative data generated by the numerous researchers and research communities fare very poorly along the FAIR Data Principles. Indeed, secondary data are hardly findable and accessible, not to speak of interoperable. Tertiary data such as descriptive statistics may be found and accessed in the scholarly literature. Being printed in PDF documents, they are, however, hardly interoperable. As a result, secondary data generated by the numerous research groups and researchers in the community cannot be reused (e.g. integrated); information systems underperform in search, retrieval or processing of tertiary data; and the reproducibility of tertiary data is generally impossible.

The New Particle Formation Event Analysis VRE[5] prototyped how infrastructure can ensure FAIR secondary and tertiary data by design. FAIRification is built into the

[3] https://www.atm.helsinki.fi/SMEAR/.

[4] https://www.go-fair.org/fields-of-action/go-build/fairification-process/.

[5] https://marketplace.eosc-portal.eu/services/new-particle-formation-event-analysis.

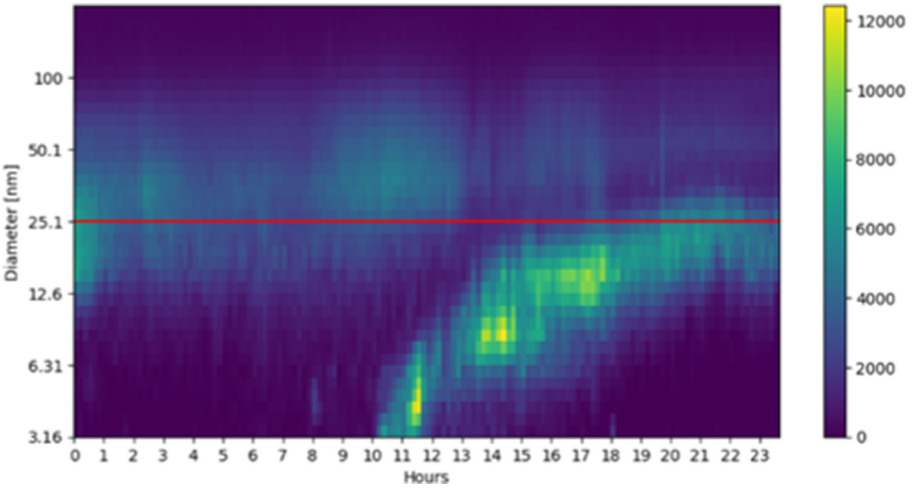

Fig. 6. Visualization of observational data for specific spatio-temporal locations. The data product shows an NPF event starting at approximately 10 am and ending shortly before 12 pm (noon). The high concentration (yellow) of particles with initially small but growing diameter size forms the typical "fingerprint" of an NPF event in observational data. (Color figure online)

infrastructure which thus ensures that data are born FAIR and frees researchers or data curators from having to FAIRify data retrospectively. Most importantly, we move data analysis into the VRE and thus harmonise data analysis across research groups; systematically catalogue secondary and tertiary data to ensure their findability and accessibility; and use languages for knowledge representation to ensure data interoperability.

4.1 Virtual Research Environment

Building on D4Science and EGI Jupyter e-Infrastructures, we developed a VRE that demonstrates how the NPF research community could perform event classification and statistical computation while the infrastructure ensures FAIR derivative (secondary and tertiary) data.

Figure 7 illustrates the VRE system architecture, its components and interactions. The NPF research community, its research groups and individual researchers access the VRE via D4Science authentication and authorization. In addition to standard VRE functionality, e.g. document management, this VRE leverages EGI Jupyter and D4Science Data Miner to provide an NPF data analysis environment with FAIR derivative data. Specialised Python functions backed by Data Miner algorithm implementations support the following operations:

- Via SmartSMEAR API, fetch primary data published by the SMEAR research infrastructure;
- Plot primary data to generate and visualise the data product required to determine whether or not an event occurred at a specified spatio-temporal location (Fig. 6);

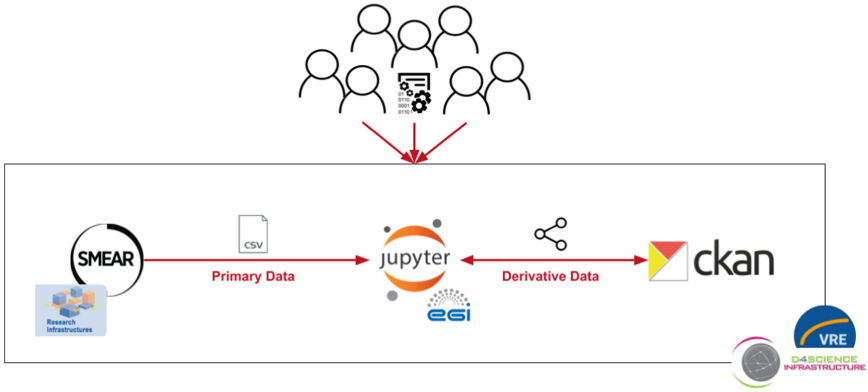

Fig. 7. VRE system architecture, its components and interactions.

- Using languages for knowledge representation (specifically, RDFS and OWL), represent derivative data (e.g. event descriptions with their attributes) richly described with a plurality of accurate and relevant attributes using vocabularies that meet domain-relevant community standards and follow FAIR principles (e.g. http://purl.obolibrary.org/obo/ENVO_01001359);
- Catalogue derivative data using the CKAN powered D4Science Catalogue;
- Retrieve catalogued secondary data (i.e., event descriptions) for statistical processing.

4.2 Benefits, Limitations and Challenges

The key benefit of the VRE is that by sharing a well-engineered computational environment for NPF event classification and statistical analysis, the research community produces FAIR derivative data without giving it any thought. In contrast to the current practice in this research community where derivative (in particular secondary) data are of high syntactic and semantic heterogeneity and impossible to integrate easily, in the VRE derivative data are automatically identified and catalogued, and thus meet key data findability and accessibility principles. Furthermore, by using languages for knowledge representation and a plurality of descriptive attributes according to domain-relevant FAIR vocabularies, derivative data also meet key data interoperability and reusability principles. Since these features are built into the infrastructure, they appear invisible to the individual researcher who can thus focus on data analysis without being exposed to the complexity of data FAIRification.

A second benefit is that derivative data are FAIR at birth rather than FAIRified retrospectively, e.g. by data curators of a research infrastructure data centre or a data publisher. FAIRification is a complex process that requires considerable domain expertise and often relies on tacit information known only to the researcher. FAIRifying early rather than later makes good sense and ensuring data are FAIR at birth is arguably the most attractive option.

A third benefit is that the VRE eliminates the need to download and upload data to and from a local computing environment (e.g. a workstation). The specialised Python

functions ensure that primary data are fetched via the SmartSMEAR API and read into native Python data structures (e.g. data frames) to enable arbitrary data processing in Jupyter. Similarly, derivative data are automatically catalogued and can be retrieved into native Python data structures from the catalogue.

A fourth benefit is that individual researchers and research groups in the community can potentially collaborate on program code development and easily share a common code base, rather than implementing scripts individually. This increases efficiency and likely software quality. Furthermore, it is trivial to add a new member of a research group (e.g. a new PhD student) to such an environment. The new member can readily benefit from work done by her colleagues, potentially even from the larger research community.

While the approach has a number of important benefits that contribute significantly to FAIR research data as well as reproducibility in science, there exist limitations. First, the development of such kind of VREs is very resource-intensive. While efficiency gains may be possible, e.g. by factoring out and reusing components that are commonly required by such VREs, research data analysis is highly contextual and difficult to generalise (and thus scale) without losing efficacy. While e-Infrastructure service providers could develop services for commonly required functionality, the development of specialised (Python) scripts for data analysis in Jupyter must rely on contributions from the research community.

The development of vocabularies that meet domain-relevant community standards and follow FAIR principles is equally resourced intensive and typically relies on strong ICT specialists and research community co-development. Researchers are mostly unaware of the benefits of such vocabularies for data (machine-to-machine) interoperability and even if the benefits are acknowledged the significant resources required to develop such vocabularies compete with research activities, which (arguably rightly so) are always prioritised over good research data management.

A relatively minor technical limitation is the poor performance of retrieving data from the catalogue. While the catalogue may be an approach to deposit data, it is not ideal for fast retrieval of data needed in the analysis. To address this performance issue, the VRE system architecture should employ more efficient intermediate data storage systems.

The challenges are perhaps more important than the current limitations. The pressing challenges of the presented VRE-based approach to NPF analysis are predominantly social. A particularly pressing one is how to motivate individual researchers to use the VRE instead of their local computational environments. Probably the most important barrier to adoption by the research community is the maturity of data analysis program code. The most advanced researchers in this community have developed mature scripts that precisely serve their needs. The key objection from such researchers is thus the maturity of the code served in the VRE. Addressing this objection is non-trivial because code maturity naturally relies virtually entirely on contributions from the research community.

Furthermore, the automated cataloguing of research data on e-Infrastructure, often perceived as potentially beyond the control of the individual researcher who created the data, is an additional barrier to adoption. Trust in e-Infrastructures that the data are safe and embargoed until at least publication is not a given but must be earned. Unfortunately, trust is gained largely through experience with working with e-Infrastructures and the

kind of VREs described here - an experience which, unfortunately, most researchers are unlikely to gain easily.

5 Mosquito Diseases Study

This science demonstrator illustrates how a LifeWatch researcher can easily upload and integrate an R-based algorithm in D4Science, making it available to other researches, in particular members of the VRE in which the algorithm was published. Once published, researchers can discover the algorithm and use it with their own data. It is also possible to adapt the algorithm and to share improved versions. When processing data-intensive analysis algorithms, the computation can be outsourced on federated resources, such as those provided by the EGI e-Infrastructures.

The scientific vision of this science demonstrator is to enable more efficient management of mosquito-borne diseases and nuisance mosquitoes. Mosquito-borne infections are among the most important new and emerging diseases globally and in Europe, and in order to predict diseases transmission areas, statistical correlation approaches are used.

LifeWatch RI provides advanced ICT, such as BioVel, supporting biodiversity research. However, it currently only provides standard algorithms for data processing. There is a need to support individual researchers' requests, e.g. import a new set of hydrological data layers into the analysis, add new algorithms that handle presence/absence into analysis etc., and a need for access to Cloud resources, e.g. to execute a large number of analytical cycles for many species under different climate scenarios.

These objectives should be achieved following the technical vision of supporting researchers in combining biological and hydrological data in a collaborative and evolving Virtual Research Environment (VRE) allowing intensive statistical computations: researchers should be able to easily share and use algorithms that they can adapt and use with their own data.

5.1 Architecture

The proposed service architecture is shown in Fig. 8. It combines different infrastructures: at a lower layer is the LifeWatch RI, containing the Swedish LifeWatch Portal that provides high-quality biological data for mosquito species, and the community data repositories that preserve environmental information and a series of ecological modelling algorithms. Datasets to be exploited include species data (95,730 abundance measurements from Sweden, Denmark, and Germany for 40 disease-carrying species in 2016), and hydrological data (generated by a regional hydrological model using 15 land-use types and 8 soil types).

At the middle layer is the EGI e-infrastructure, which provides cloud computation and storage resources supporting data-intensive workflow executions.

At the top layer is the D4Science VRE and the Biodiversity Virtual e-Laboratory (BioVel) portal, that provide high-level user interfaces. BioVel[6] is a software environment that assists scientists in collecting, organising, and sharing data processing and analysis

[6] https://www.biovel.eu/.

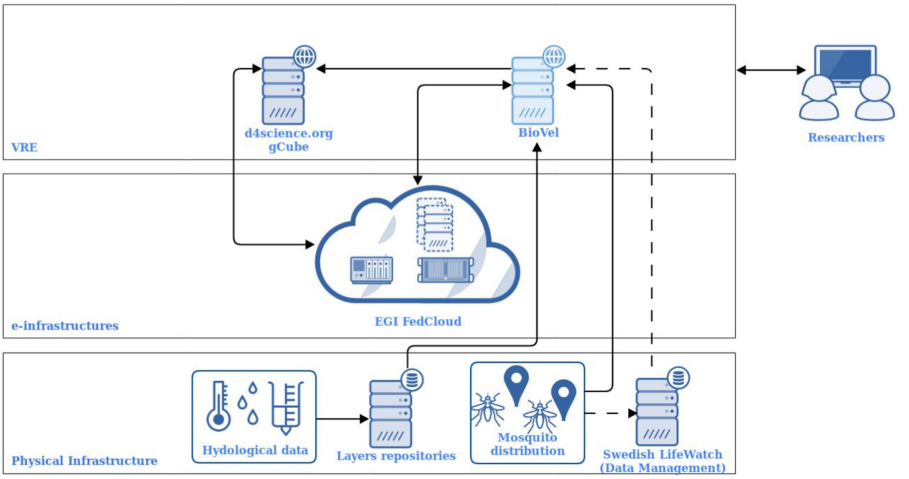

Fig. 8. Architecture includes three layers: 1) Physical Infrastructure, 2) e-Infrastructure, and 3) VRE.

tasks in biodiversity and ecological research. The service components of the platform include a Biodiversity Catalogue (a library with well-annotated data and analysis services), the data processing environments (such as RStudio for creating R programs), a workbench (for assembling data access and analysis pipelines), the myExperiment workflow library (that stores existing workflows), and the BioVel Portal (that allows researchers and collaborators to execute and share workflows).

The existing BioVel platform can generate environmental values from species occurrences, however, it only provides standard analysis algorithms. Integrating the D4Science and gCube -based VRE can enrich the functionality of the LifeWatch ICT to allow dynamic modelling.

5.2 User Interface

The D4Science/gCube-based VRE for mosquito disease study has been set up with the support from T7.1. The interfaces for the mosquito disease study are shown in Fig. 9, Fig. 10. It provides a programming environment (shown in Fig. 10), and it allows biodiversity researchers to develop and compile own/customised analysis algorithms using R, CLI, etc. A researcher can decide to share his/her data, algorithms, or workflows by publishing them in the group area (shown in Fig. 9) that enables social communications via messages, comments, etc.

5.3 Advantages

Using the VRE, there is no more need for manual sharing of data and algorithms. Information is always synchronised, and data and algorithms are joint in a single place. Users can enjoy an easy and user-friendly access interface. The D4Science/gCube-based VRE has an interface to EGI Cloud/HTC resources. If needed, it can outsource the

Fig. 9. VRE area for sharing data, algorithms, and workflows.

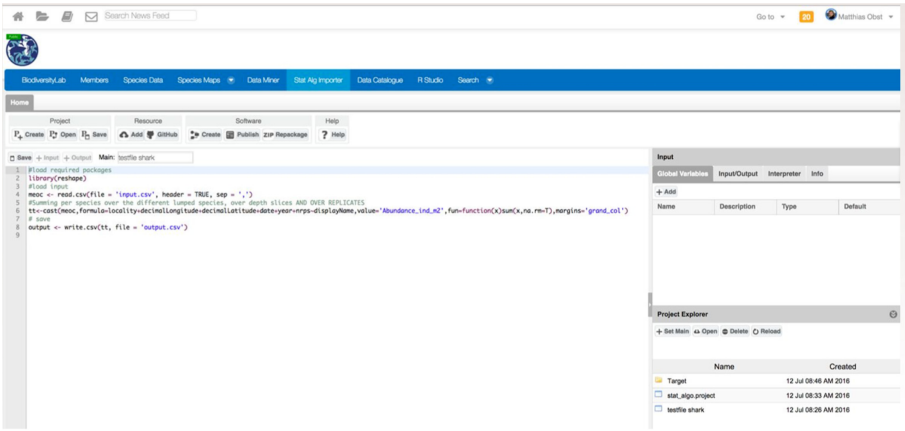

Fig. 10. VRE area for developing an analysis algorithm.

computation on the large-scale e-Infrastructure that can handle computation in parallel and store and share large volumes of data.

The integration service can bring added value to the Lifewatch community. It makes it possible for individual researchers to repeat and reuse algorithms at will, run trend analysis, and add new parameters and custom data. The VRE provides provenance registration that improves reproducibility. The VRE also allows retention of computation results in the user's workspace. This makes it possible to edit and adapt algorithms.

The integration service also brings added value to ENVRIplus community. Enabling individual researchers to share data and/or algorithms is common to many ENVRIplus RIs where currently data is processed using standard models. Researchers want to use different analysis models and they need a VRE to work together.

This pilot investigation tested and validated WP7 technology. The demo illustrates the integration solutions of linking gCube VRE to LifeWatch RI and to the EGI e-Infrastructure. There are also some lessons learned from the pilot activities: The D4Science/gCube VRE is easy for simple algorithms

. It needs integration efforts for complicated algorithms that request domain researchers to have technical skills to work with different techniques.

6 Conclusion

This chapter presented several diverse science demonstrators that were implemented by building on state-of-the-art D4Science e-Infrastructure to realise specific VREs. The demonstrators show that D4Science is capable of supporting the implementation of complex data processing and analysis pipelines and, more importantly, does so efficiently by ensuring the reuse of services and support extensions to VRE functionality with user-defined functions (scripts). The strong encapsulation of user-defined functions in D4Science (in contrast to, e.g. Jupyter notebooks) can at first be seen as an unwanted overhead but comes with advantages. First, the functions are automatically exposed as Web Processing Services and can be called also from third-party systems (e.g. Taverna workflows). Second, being a collaborative environment, D4Science ensures that collaborators do not inadvertently modify processing and thus potentially introduce errors. Moving individual researchers and entire research communities from their local computing environments into VREs is surely a monumental task in its own right. However, there are a lot of arguments for it, one being that infrastructures and communities of practice can ensure that research data are born FAIR instead of being FAIRyfied in a subsequent stage.

Acknowledgements. This work was supported by the European Union's Horizon 2020 research and innovation programme via the ENVRIplus project under grant agreement No. 654182.

References

1. Asmi, A., Kutsch, W.L.: ENVRI PLUS: European initiative towards technical and research cultural solutions for across-disciplines accessible Research Infrastructure products. AGU Fall Meeting Abstracts, IN31B-1764 (2015)
2. Assante, M., et al.: The gCube system: delivering virtual research environments as-a-service. Future Gener. Comput. Syst. **95**, 445–453 (2019). https://doi.org/10.1016/j.future.2018.10.035
3. Assante, M., et al.: Enacting open science by d4science. Future Gener. Comput. Syst. **101**, 555–563 (2019). https://doi.org/10.1016/j.future.2019.05.063
4. Aubinet, M., Vesala, T., Papale, D.: Eddy Covariance: A Practical Guide to Measurement and Data Analysis. Springer, heidelberg (2012). https://doi.org/10.1007/978-94-007-2351-1
5. Barker, M., et al.: The global impact of science gateways, virtual research environments and virtual laboratories. Future Gener. Comput. Syst. **95**, 240–248 (2019). https://doi.org/10.1016/j.future.2018.12.026
6. Bartling, S., Friesike, S. (eds.): Opening Science. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-00026-8
7. Dada, L., et al.: Refined classification and characterization of atmospheric new-particle formation events using air ions. Atmos. Chem. Phys. **18**(24), 17883–17893 (2018). https://doi.org/10.5194/acp-18-17883-2018

8. Fratini, G., Mauder, M.: Towards a consistent eddy-covariance processing: an intercomparison of EddyPro and TK3. Atmos. Meas. Tech. **7**(7), 2273–2281 (2014). https://doi.org/10.5194/amt-7-2273-2014
9. Hellström, M., et al.: Near real time data processing In: ICOS RI. In Proceedings of 2nd International Workshop on Interoperable Infrastructures for Interdisciplinary Big Data Sciences (IT4RIs 16), Porto, Portugal, vol. 30, November 2016
10. Jones, S., Abramatic, J.-F. (eds.): European Open Science Cloud (EOSC) Strategic Implementation Plan. European Commission (2019). https://doi.org/10.2777/202370
11. Sabbatini, S.: Eddy covariance raw data processing for CO2 and energy fluxes calculation at ICOS ecosystem stations. Int. Agrophys. (2018). https://doi.org/10.1515/intag-2017-0043
12. Wilczak, J.M., Oncley, S.P., Stage, S.A.: Sonic anemometer tilt correction algorithms. Bound.-Layer Meteorol. **99**(1), 127–150 (2001). https://doi.org/10.1023/A:1018966204465